

This Page Is Inserted by IFW Operations
and is not a part of the Official Record

BEST AVAILABLE IMAGES

Defective images within this document are accurate representations of the original documents submitted by the applicant.

Defects in the images may include (but are not limited to):

- BLACK BORDERS
- TEXT CUT OFF AT TOP, BOTTOM OR SIDES
- FADED TEXT
- ILLEGIBLE TEXT
- SKEWED/SLANTED IMAGES
- COLORED PHOTOS
- BLACK OR VERY BLACK AND WHITE DARK PHOTOS
- GRAY SCALE DOCUMENTS

IMAGES ARE BEST AVAILABLE COPY.

**As rescanning documents *will not* correct images,
please do not report the images to the
Image Problems Mailbox.**

(19) 日本国特許庁 (J P)

(12) 公開特許公報 (A)

(11) 特許出願公開番号

特開平10-198642

(43) 公開日 平成10年(1998) 7月31日

(51) Int.Cl.⁶

G 0 6 F 15/16
13/00

識別記号

3 7 0
3 5 7

F I

G 0 6 F 15/16
13/00

3 7 0 N
3 5 7 Z

審査請求 未請求 請求項の数9 O L (全 14 頁)

(21) 出願番号

特願平9-1970

(22) 出願日

平成9年(1997) 1月9日

(71) 出願人 000005223

富士通株式会社

神奈川県川崎市中原区上小田中4丁目1番
1号

(72) 発明者 大橋 勝之

神奈川県川崎市中原区上小田中4丁目1番
1号 富士通株式会社内

(74) 代理人 弁理士 山谷 皓榮 (外2名)

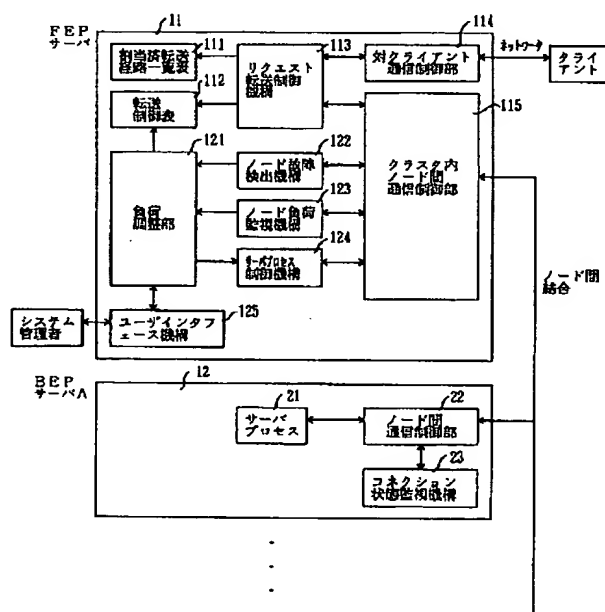
(54) 【発明の名称】 サーバ装置

(57) 【要約】

【課題】 1種類のサービスの処理を複数台のバックエンドサーバで負荷分散して処理性能を向上すること。

【解決手段】 クライアントのリクエストを処理するサーバプロセス21を配置した複数個のバックエンドサーバ12と、クライアントからのリクエストを受信するフロントエンドサーバ11と、該フロントエンドサーバ11に配置し、前記受信したリクエストを適切な前記バックエンドサーバ12に転送するリクエスト転送制御機構113とを備え、該リクエスト転送制御機構113は、クライアントの識別情報を用いて、同一サービスに対する複数のクライアントからのリクエストを転送する前記バックエンドサーバ12を決定してリクエストを転送する。

本発明の原理説明図



【特許請求の範囲】

【請求項1】クライアントのリクエストを処理するサーバプロセスを配置した複数のバックエンドサーバと、クライアントからのリクエストを受信するフロントエンドサーバと、

該フロントエンドサーバに配置し、前記受信したリクエストを適切な前記バックエンドサーバに転送するリクエスト転送制御機構とを備え、

該リクエスト転送制御機構は、クライアントの識別情報を用いて、同一サービスに対する複数のクライアントからのリクエストを転送する前記バックエンドサーバを決定してリクエストを転送することを特徴としたサーバ装置。

【請求項2】前記リクエスト転送制御機構は、リクエスト転送制御表を用いて転送するリクエストの比率をバックエンドサーバ毎に制御することを特徴とした請求項1記載のサーバ装置。

【請求項3】サーバの故障を検出するノード故障検出機構を設け、

該ノード故障検出機構がサーバの故障を検出すると、リクエスト転送制御機構は、故障したバックエンドサーバへのリクエスト転送を中止して、クライアントからの再リクエスト時に正常運用している別のバックエンドサーバにリクエストを転送することを特徴とした請求項1又は2記載のサーバ装置。

【請求項4】前記バックエンドサーバの負荷を監視してリクエストの転送比率を変更するノード負荷監視機構を設け、

該ノード負荷監視機構が負荷が高い前記バックエンドサーバを発見すると、リクエスト転送制御機構は、負荷の高いバックエンドサーバへの転送比率を下げて、前記バックエンドサーバ間の負荷を平均化することを特徴とした請求項1～3のいずれかに記載のサーバ装置。

【請求項5】システム管理者が前記バックエンドサーバへの転送比率を制御するためのユーザインタフェース機構を設けることを特徴とした請求項1～4のいずれかに記載のサーバ装置。

【請求項6】前記サーバプロセスとクライアントプロセス間の接続状態を監視する接続状態監視機構を設け、

前記リクエスト転送制御機構は、前記接続状態監視機構に接続状態を問い合わせ、サービス中の接続に対しては、該サービス終了後に、転送経路の変更を行うことを特徴とした請求項1～5のいずれかに記載のサーバ装置。

【請求項7】前記サーバプロセスの制御を行うサーバプロセス制御機構を設け、

あるサービスのサーバプロセスが配置された各バックエンドサーバの負荷が高くなった時に、前記サーバプロセス制御機構はそのサーバプロセスが未配置なバックエン

ドサーバにサーバプロセスを起動することを特徴とした請求項1～6のいずれかに記載のサーバ装置。

【請求項8】前記リクエスト転送制御機構を、ネットワークの通信制御をするネットワークドライバとパケットの処理をするオペレーティングシステムのパケット処理部の中間に配置したパケットフィルタで構成することを特徴とした請求項1～7のいずれかに記載のサーバ装置。

【請求項9】前記クライアントの識別情報として、クライアントから受信したリクエストのパケットのソースアドレスとソースポート番号のペアを使用することを特徴とした請求項1～8のいずれかに記載のサーバ装置。

【発明の詳細な説明】

【0001】

【発明の属する技術分野】本発明は、計算機環境におけるクライアント／サーバシステムのサーバの構成に関し、特に、サーバをクラスタシステム（複数のノードを結合したシステム）で構成する際のノード（サーバ）間の負荷分散を行うサーバ装置に関する。

【0002】近年、インターネット等の開放型システムの普及に伴い、クライアント／サーバシステムへの関心が高まり、特に不特定多数のクライアントからの要求が集中することが予想されるサーバシステムを高性能／高可用化する（利用可能率を高くする）ための技術として、クラスタシステムの負荷分散に関する技術の開発が望まれている。

【0003】

【従来の技術】従来のクラスタシステムにおける負荷分散技術の1つに、ルータやパケットフィルタによるリダイレクション（仕事の割り振り）がある。この方式では、サーバはクライアントからのリクエストを受けるフロントエンドノード（FEPノード）と、クライアントからのリクエストを処理する機能を複数台のノードに分散配置したバックエンドノード（BEPノード）群から構成されていた。全てのクライアントはFEPノードに対してリクエストを送り、FEPノードは要求されたサービスに対して適切なBEPノードにリクエストを転送して負荷分散を行っていた。

【0004】図7は従来例の説明図（1）、図8は従来例の説明図（2）である。以下図7、図8に基づいて従来例の説明をする。図7において、従来のリダイレクションを利用した負荷分散の例であり、サーバドメイン1と複数のクライアントドメイン、がWAN（ワイドエリアネットワーク）を介して接続されている。

【0005】サーバドメイン1には、FEPサーバ11、複数のサーバ12（サーバA～C）、ドメインネームサーバ（DNS）13、ルータ14が設けてある。クライアントドメイン1には、複数のクライアント（1）～（3）、DNS43、ルータ44が設けてある。クライアントドメイン2には、複数のクライアント（4）～

(6)、DNS43、ルータ44が設けてある。

【0006】FEPサーバ11は、LAN（ローカルエリアネットワーク）でクライアントからのリクエストを受信し、高速ノード間結合により適切なBEPノードにリクエストを転送する転送制御プロセスを有するFEPノードである。サーバ12（サーバA～C）は、それぞれ異なる種類のサービスX～Zを処理するサーバプロセスX～Zを有するBEPノードである。

【0007】DNS13は、LANと接続されておりクライアント等から問い合わせのあったサービス名に対応するIPアドレス（各コンピュータ固有のアドレス）を応答するものである。例えば、図8（b）はIPアドレス例の説明であり、図8（b）のようにクライアントは、DNS13からサービス名「ftp-server」のIPアドレス「133.160.12.5」を取得して、この取得したIPアドレスに向かってリクエストを出すものである。ルータ14は、WANとLAN間の中継処理を行うものである。

【0008】複数のクライアント（1）～（3）、複数のクライアント（4）～（6）は、それぞれのLANで接続されておりサービス（データ処理）を要求する側である。DNS43は、LANと接続されており問い合わせのあったサービス名に対応するIPアドレス応答するものである。ルータ44は、LANとWAN間の中継処理を行うものである。

【0009】図8（a）は仕事の割り振り表の説明であり、FEPノード上のプロセスによって管理される各サーバプロセスの配置を示すものである。図8（a）において、サービス識別子に対応する転送先サーバ識別子が示されている。例えば、サービスXはサーバAに、サービスYはサーバBに、サービスZはサーバCに転送することが示されている。

【0010】全てのクライアントからのリクエストは、クライアントドメインのLAN、ルータ44、WAN、ルータ14、サーバドメイン1のLANを介してFEPノードに送られ、FEPノードは、要求されたサービスの種類に応じてリクエストの転送先ノードを決定し、適切なBEPノードに、高速ノード間結合を介してリクエストを転送する。BEPノードA～C上のサーバプロセスX～Zは、FEPノードから転送されたリクエストを処理してクライアントにリプライするものであった。

【0011】

【発明が解決しようとする課題】前記従来のリダイレクション方式においては、BEPノード毎に異なるサービスを処理するサーバプロセスを配置し、図8（a）のような表を用いて、クライアントが要求したサービスをキーとして転送先BEPノードを決定していた。このため、1種類のサービスの処理が多くなっても複数台のBEPノードで負荷分散することができない課題があった。

【0012】本発明は、このような従来の課題を解決し、同一サービスに対する複数のクライアントからのリクエストを複数台のサーバノードで負荷分散をできるようにし、また、自動的にサーバノード故障隠蔽機能や、負荷調整機能を持つことにより、システム管理コストを削減することを目的とする。

【0013】

【課題を解決するための手段】図1は本発明の原理説明図である。図1中、11はフロントエンドサーバ、12はバックエンドサーバ、21はサーバプロセス、22はノード間通信制御部、23はコネクション状態監視機構、111は割当済転送経路一覧表、112はリクエスト転送制御表、113はリクエスト転送制御機構、114は対クライアント通信制御部、115はクラスタ内ノード間通信制御部、121は負荷調整部、122はノード故障検出機構、123はノード負荷監視機構、124はサーバプロセス制御機構、125はユーザインタフェース機構である。

【0014】本発明は前記従来の課題を解決するため次のように構成した。

（1）：クライアントのリクエストを処理するサーバプロセス21を配置した複数個のバックエンドサーバ12と、クライアントからのリクエストを受信するフロントエンドサーバ11と、該フロントエンドサーバ11に配置し、前記受信したリクエストを適切な前記バックエンドサーバ12に転送するリクエスト転送制御機構113とを備え、該リクエスト転送制御機構113は、クライアントの識別情報を用いて、同一サービスに対する複数のクライアントからのリクエストを転送する前記バックエンドサーバ12を決定してリクエストを転送する。

【0015】（2）：前記（1）のサーバ装置において、前記リクエスト転送制御機構113は、リクエスト転送制御表112を用いて転送するリクエストの比率をバックエンドサーバ12毎に制御する。

【0016】（3）：前記（1）又は（2）のサーバ装置において、サーバの故障を検出するノード故障検出機構122を設け、ノード故障検出機構122がサーバの故障を検出すると、リクエスト転送制御機構113は、故障したバックエンドサーバ12へのリクエスト転送を中止して、クライアントからの再リクエスト時に正常運用している別のバックエンドサーバ12にリクエストを転送する。

【0017】（4）：前記（1）～（3）のサーバ装置において、前記バックエンドサーバ12の負荷を監視してリクエストの転送比率を変更するノード負荷監視機構123を設け、該ノード負荷監視機構123が負荷が高い前記バックエンドサーバ12を発見すると、リクエスト転送制御機構113は、負荷の高いバックエンドサーバ12への転送比率を下げて、前記バックエンドサーバ12間の負荷を平均化する。

【0018】(5):前記(1)～(4)のサーバ装置において、システム管理者が前記バックエンドサーバ12への転送比率を制御するためのユーザインタフェース機構125を設ける。

【0019】(6):前記(1)～(5)のサーバ装置において、前記サーバプロセス21とクライアントプロセス間のコネクション状態を監視するコネクション状態監視機構23を設け、前記リクエスト転送制御機構113は、前記コネクション状態監視機構23にコネクション状態を問い合わせ、サービス中のコネクションに対しては、該サービス終了後に、転送経路の変更を行う。

【0020】(7):前記(1)～(6)のサーバ装置において、前記サーバプロセス21の制御を行うサーバプロセス制御機構124を設け、あるサービスのサーバプロセス21が配置された各バックエンドサーバ12の負荷が高くなった時に、前記サーバプロセス制御機構124がそのサーバプロセス21が未配置なバックエンドサーバ12にサーバプロセス21を起動する。

【0021】(8):前記(1)～(7)のサーバ装置において、前記リクエスト転送制御機構113を、ネットワークの通信制御をするネットワークドライバとパケットの処理をするオペレーティングシステムのパケット処理部の間に配置したパケットフィルタで構成する。

【0022】(9):前記(1)～(8)のサーバ装置において、前記クライアントの識別情報として、クライアントから受信したリクエストのパケットのソースアドレスとソースポート番号のペアを使用する。

【0023】(作用)前記構成に基づく作用を説明する。複数のバックエンドサーバ12に配置したサーバプロセス21でクライアントのリクエストを処理し、フロントエンドサーバ11に配置したリクエスト転送制御機構113で受信したクライアントからのリクエストをクライアントの識別情報を用いて、同一サービスに対する複数のクライアントからのリクエストを転送する前記バックエンドサーバ12を決定してリクエストを転送する。このため、1種類のサービスの処理を複数のバックエンドサーバで負荷分散でき、処理性能を向上することができる。

【0024】また、前記リクエスト転送制御機構113で、リクエスト転送制御表112を用いて転送するリクエストの比率をバックエンドサーバ12毎に制御する。このため、バックエンドサーバの処理能力に応じた負荷分散ができ、より処理性能を向上することができる。

【0025】さらに、ノード故障検出機構122がサーバの故障を検出すると、リクエスト転送制御機構113で、故障したバックエンドサーバ12へのリクエスト転送を中止して、クライアントからの再リクエスト時に正常運用している別のバックエンドサーバ12にリクエストを転送する。このため、クライアントからの再リクエストにより、バックエンドサーバの故障をクライアント

から隠蔽することができる。

【0026】また、ノード負荷監視機構123が負荷が高い前記バックエンドサーバ12を発見すると、リクエスト転送制御機構113で、負荷の高いバックエンドサーバ12への転送比率を下げて、前記バックエンドサーバ12間の負荷を平均化する。このため、常に負荷を平均化することができ、処理性能を向上することができる。

【0027】さらに、ユーザインタフェース機構125で、システム管理者が前記バックエンドサーバ12への前記転送比率を制御する。このため、例えば、活性保守のためのバックエンドサーバの転送比率をゼロとしてノードの一時的な切り離しや再組み込みが可能となる。

【0028】また、前記リクエスト転送制御機構113は、コネクション状態監視機構23にコネクション状態を問い合わせ、サービス中のコネクションに対しては、該サービス終了後に、転送経路の変更を行う。このため、サービス途中に新たなバックエンドサーバに切り換えることがなく、処理の中断を防止することができる。

【0029】さらに、あるサービスのサーバプロセス21が配置された各バックエンドサーバ12の負荷が高くなった時に、サーバプロセス制御機構124がそのサーバプロセス21が未配置なバックエンドサーバ12にサーバプロセス21を起動する。このため、複数のバックエンドサーバ全体の負荷を調整することができ、処理性能を向上することができる。

【0030】また、前記リクエスト転送制御機構113を、ネットワークの通信制御をするネットワークドライバとパケットの処理をするオペレーティングシステムのパケット処理部の間に配置したパケットフィルタで構成する。このため、パケットフィルタで1種類のサービスの処理を複数のバックエンドサーバで負荷分散でき、処理性能が向上する。

【0031】さらに、前記クライアントの識別情報として、クライアントから受信したリクエストのパケットのソースアドレスとソースポート番号のペアを使用する。このため、クライアントプロセス毎にリクエストの転送先ノードの制御を行うことができる。

【0032】

【発明の実施の形態】図2～図6は本発明の実施の形態を示した図である。本発明では、FEPノードにおけるリクエスト転送制御に、クライアントが要求したサービスの種類に加えて、リクエストを送ったクライアント識別子を用い、同一サービスに対するリクエストの転送先BEPノードをクライアント毎に制御するものである。以下、図面に基づいて本発明の実施の形態を説明する。

【0033】(1):サーバ装置の説明

:装置構成の説明

図2は装置構成図(1)である。以下、図2に基づいてサーバ装置の説明をする。

【0034】図2において、サーバ装置は、FEPサーバ（FEPノード）11と複数のBEPサーバ（BEPノード）12a、12b、12c、・・・がノード間結合3で接続されている。クライアント4は、LANあるいはWAN5を経由してFEPノード11と結合されている。

【0035】FEPサーバ11には、割当済転送経路一覧表111、転送制御表112、リクエスト転送処理部113、対クライアント通信制御部114、クラスタ内ノード間通信制御部115、負荷調整部121、ノード故障検出部122、ノード負荷監視部123、サーバプロセス制御部124、システム管理者インタフェース125が設けてある。

【0036】BEPサーバ12aには、サーバプロセス21a、ノード間通信制御部22a、コネクション状態監視部23aが設けてあり、BEPサーバ12bには、サーバプロセス21b、ノード間通信制御部22b、コネクション状態監視部23bが設けてあり、BEPサーバ12cにも、サーバプロセス21c、ノード間通信制御部22c、コネクション状態監視部23cが設けてある。

【0037】割当済転送経路一覧表111は、クライアント毎に割り当てたリクエストの転送経路を記録したものである。転送制御表112は、リクエストを処理するサーバプロセスの配置とリクエストの振り分け比率を記録したものである。リクエスト転送処理部113は、クライアントからのリクエストをBEPノードに転送するものである。対クライアント通信制御部114は、クライアントとの通信を制御するものである。クラスタ内ノード間通信制御部115は、BEPノードとの通信を制御するものである。負荷調整部121は、システム管理者の要求や発生したイベントに応じて動的に転送制御を調整するものである。ノード故障検出部122は、BEPノードの故障を検出するものである。ノード負荷監視部123は、BEPノードの負荷状態を監視するものである。サーバプロセス制御部124は、BEPノードに必要なサーバプロセスを起動するものである。システム管理者インタフェース125は、システム管理者6にユーザインタフェースを提供するものである。

【0038】サーバプロセス21a、21b、21cは、クライアントからのリクエストを処理するプロセスである。ノード間通信制御部22a、22b、22cは、FEPノードとの通信を制御するものである。コネクション状態監視部23a、23b、23cは、クライアントとサーバプロセスとのコネクション状態を監視するものである。

【0039】：転送制御表と割当済転送経路一覧表の説明

図3は転送制御表と割当済転送経路一覧表の説明図であり、図3（a）は転送制御表の説明、図3（b）は割当

済転送経路一覧表の説明である。

【0040】図3（a）において、転送制御表112には、リクエストの宛て先としてのポート番号等を使用するサービス識別子、転送先となる利用可能なBEPノードのIPアドレス等を使用する利用可能BEPノード識別子、同一サービスに対するリクエストの分配比率（転送比率）である比率が設けてある。

【0041】図3（b）において、割当済転送経路一覧表111には、リクエストの宛て先としてのポート番号等を使用するサービス識別子、ソースIPアドレスとソースポートのポート番号等を使用するクライアント識別子、転送先のBEPノードのIPアドレス等を使用する転送先BEPノード識別子が設けてある。

【0042】（2）：サーバ装置の動作の説明

以下、図2に基づいてサーバ装置の動作を説明する。

：クライアントからリクエストがあった場合の説明
クライアント4がリクエストをLANあるいはWAN5を経由してFEPノード11に送ると、対クライアント通信制御部114によって受信されたリクエストはリクエスト転送処理部113に送られる。リクエスト転送処理部113は、まず、割当済転送経路一覧表111を用いて、すでに転送経路を割り当てたことがあるクライアントからのリクエストかどうか調べる。この割当済転送経路一覧表111には、リクエストしたクライアントの識別子と要求したサービスの識別子のペアをキーとして、割り当てられた転送先BEPノードの識別子が記録されている。

【0043】割当済転送経路一覧表111にクライアントのエントリがない場合には、リクエスト転送処理部113は、転送制御表112を参照して、転送先BEPノードを決定する。この転送制御表112には、利用可能BEPノード識別子とリクエストの分配比率がサービス毎に記録されている。

【0044】転送制御表112でリクエストの転送先BEPノードが決定すると、リクエスト転送処理部113は、割当済転送経路一覧表111に転送先を割り当てたクライアントのエントリを作り、クラスタ内ノード間通信制御部115、ノード間結合3を経由して、リクエストを転送先として決定したBEPノードに転送する。

【0045】このリクエストが転送されたBEPノード（例えば、BEPノード12aとする）では、ノード間通信制御部22aがリクエストを受信してサーバプロセス21aにリクエストを送る。サーバプロセス21aは、リクエストを処理すると、ノード間通信制御部22a、ノード間結合3を経由して、FEPノード11にリプライを送る。FEPノード11は、クラスタ内ノード間通信制御部115でリプライを受け、リクエスト転送処理部113が対クライアント通信制御部114、LANあるいはWAN5を介してクライアント4にリプライを返す。

【0046】割当済転送経路一覧表111にエントリが存在するクライアントからのリクエストの場合は、リクエスト転送処理部113は、割当済転送経路一覧表111で割り当てられているBEPノードにリクエストを送ることになる。このような処理を行うことによって、同一のクライアント識別子で識別されるクライアントからのリクエストは同一のBEPサーバに転送されることになる。

【0047】：ノード故障の場合の説明

ノード故障検出部122は、定期的に全てのBEPノードが正常動作していることを確認する。ノード故障検出部122は、BEPノードの故障を検出すると、負荷調整部121を経由して転送制御表112を書き換えて、故障したBEPノードへのリクエストの分配を停止する。リクエスト転送処理部113は、割当済転送経路一覧表111内の転送先BEPノードが故障したBEPノードとなっている全てのエントリを削除する。

【0048】このBEPノードの故障によって、故障したBEPノードを利用していたクライアントとBEPノード上のサーバプロセスとのコネクションは切断されるが、クライアントがリクエストをリトライすることによって故障は隠蔽される。

【0049】クライアントからの再リクエストはリクエスト転送処理部113に送られると、リクエスト転送処理部113は、割当済転送経路一覧表111を調べるが、再リクエストしたクライアントのエントリは既に削除されているため、リクエスト転送処理部113は、転送制御表112に基づいて新たな転送先を決定してリクエストを転送する。リクエストを転送されたBEPノードでは、通常のリクエストと同様にリクエストを処理してリプライする。

【0050】：ノード負荷監視の説明

ノード負荷監視部123は、定期的に全てのBEPノードの負荷状態を監視する。ノード負荷監視部123は、負荷が高いBEPノードを発見すると、負荷調整部121を経由して転送制御表112を書き換えて、負荷が高いBEPノードへのリクエストの分配比率を下げる。リクエスト転送処理部113は、分配比率変更時に割り当て済の転送経路の状態を各BEPサーバのコネクション状態監視部23a、23b、23c、・・・に問い合わせ、割当済転送経路一覧表111からアクティブでないコネクションの転送経路のエントリを削除し、その時点のリクエスト分配比率を算出する。

【0051】リクエスト転送処理部113は、以降の新規クライアントからのリクエスト転送先の調整や、コネクション状態監視部23a、23b、23c、・・・に問い合わせによる割当済転送経路一覧表111の更新により、書き換えられた転送制御表112で指定された分配比率に実際の分配比率を近付けるようにする。

【0052】：サーバプロセスの起動・停止の説明

サーバプロセス制御部124は、クラスタ内ノード間通信制御部115、ノード間結合3を介して各BEPノード上のサーバプロセスを制御するものである。

【0053】あるサービスに割り当てられた全てのBEPノードの負荷が高い場合には、ノード負荷監視部123は負荷調整部121にBEPノードの割り当て増を依頼する。各サービスへのBEPノードの割り当てを管理している負荷調整部121は、負荷が高いサービスへのBEPノード割り当ての増を決めると、新たに割り当てたBEPノードへサーバプロセスの起動をサーバプロセス制御部124に依頼する。そこで、サーバプロセス制御部124は、クラスタ内ノード間通信制御部115、ノード間結合3を経由して、指定されたBEPノード上のサーバプロセスを起動する。このサーバプロセスの起動が完了すると、負荷調整部121は、新たに割り当てたBEPノードにリクエストを転送するように転送制御表112を書き換える。

【0054】新たに割り当てたBEPノード上で、それまで動作していたサーバプロセスを停止させることが必要な場合、負荷調整部121は、転送制御表112を書き換えてリクエストに対する分配を停止する（分配比率を0とする）。リクエスト転送処理部113は、リクエストの転送の停止が指示されたBEPノードのコネクション状態監視部にコネクションの状態を問い合わせ、割当済転送経路一覧表111の未使用状態のエントリを削除する。負荷調整部121は、割当済転送経路一覧表111を参照して、指定したBEPノードへのコネクションのエントリが無くなった時点で、サーバプロセス制御部124にサーバプロセスの停止を依頼する。サーバプロセス制御部124は、クラスタ内ノード間通信制御部115、ノード間結合3を経由して、指定されたBEPノード上のサーバプロセスを停止する。

【0055】：システム管理者インタフェースの説明
システム管理者インタフェース125は、システム管理者6にシステム管理のためのインタフェースを提供するものである。

【0056】システム管理者インタフェース125が提供する機能としては、クラスタへの動的な（運用中の）ノードの追加／切り離し、サービスへのBEPノード割り当て指定／変更、その他システム状態の参照等の機能がある。

【0057】（3）：UNIX（AT&T社の汎用のオペレーティングシステム）マシンによる説明
以下、図4～図6に従ってUNIXマシンによる説明をする。

【0058】：クライアント／サーバシステムの説明
図4はクライアント／サーバシステムの説明図である。以下図4に基づいて説明をする。

【0059】図4において、リダイレクションを利用した負荷分散の例であり、サーバドメイン1と複数のクラ

イアントドメイン、がWANを介して接続されている。サーバドメイン1には、FEPサーバ11、複数のサーバ12（サーバA～F）、ドメインネームサーバ

（DNS）13、ルータ14が設けてある。クライアントドメインには、複数のクライアント（1）～

（5）、DNS43、ルータ44が設けてある。クライアントドメインには、複数のクライアント（6）～

（10）、DNS43、ルータ44が設けてある。
【0060】FEPサーバ11は、LAN（ローカルエリアネットワーク）でクライアントからのリクエストを受信し、高速ノード間結合により適切なBEPノード12にリクエストを転送するパケットフィルタ113を有するFEPノードである。サーバ12（サーバA～F）は、サービスX、Y、Zを処理するサーバプロセス21（サーバプロセスX、Y、Z）を有するBEPノードである。DNS13は、LANと接続されておりクライアント等から問い合わせのあったサービス名に対応するIPアドレス（各コンピュータ固有のアドレス）を応答するものである。

【0061】複数のクライアント（1）～（5）、複数のクライアント（6）～（10）は、それぞれのLANで接続されておりサービス（データ処理）を要求する側である。DNS43は、LANと接続されており問い合わせのあったサービス名に対応するIPアドレス応答するものである。ルータ44は、LANとWAN間の中継処理を行うものである。

【0062】：サーバ装置の説明

UNIXマシンのサーバ装置（クラスタシステム）における負荷分散について、図2のリクエスト転送処理部113は、ネットワークのデータ処理をするストリーム処理部の最下層に位置するパケットフィルタとして実現される。そして、割当済転送経路一覧表111と転送制御表112はパケットフィルタ内に置かれている。また、図2の負荷調整部121はユーザ空間内のUNIXプロセスである負荷調整プロセスとして実現する。この負荷調整プロセスは、UNIX OS（オペレーティングシステム）が提供するIOCTLシステムコールを利用して、パケットフィルタ内の転送制御表の書き換えを行うものである。

【0063】図5は装置構成図（2）である。以下、図5に基づいてサーバ装置の説明をする。図5において、サーバ装置は、FEPノード11とBEPノード12（実際は複数のBEPノードが設けられる）がノード間結合ネットワーク3で接続されている。クライアントとは、LAN等のネットワーク5を介してFEPノード11と結合されている。

【0064】FEPノード11には、カーネル（中枢部分）空間とユーザ空間があり、カーネル空間にはストリーム処理部、外部ネットワークドライバ114、内部ネットワークドライバ115が設けてある。ストリーム処

理部には、パケットフィルタ113、IPパケット処理層131、TCPパケット処理層132が設けてあり、パケットフィルタ113には、割当済転送経路一覧表111と転送制御表112が設けてある。ユーザ空間には、負荷調整プロセス121、ノード故障検出プロセス122、ノード負荷監視プロセス123、サーバ制御プロセス124、ユーザI/F処理プロセス125が設けてある。

【0065】BEPノード12にも、カーネル空間とユーザ空間があり、カーネル空間にはストリーム処理部、内部ネットワークドライバ22が設けてある。ストリーム処理部には、IPパケット処理層241、TCPパケット処理層242が設けてある。ユーザ空間には、接続監視プロセス23、サーバプロセス21が設けてある。

【0066】割当済転送経路一覧表111は、クライアント毎に割り当てたリクエストの転送経路を記録したものである。転送制御表112は、リクエストを処理するサーバプロセスの配置とリクエストの振り分け比率を記録したものである。パケットフィルタ113は、クライアントからのリクエストをBEPノードに転送するものである。外部ネットワークドライバ114は、クライアントとの通信を制御するものである。内部ネットワークドライバ115は、BEPノードとの通信を制御するものである。負荷調整プロセス121は、システム管理者の要求や発生したイベントに応じて動的に転送制御を調整するものである。ノード故障検出プロセス122は、BEPノードの故障を検出するものである。ノード負荷監視プロセス123は、BEPノードの負荷状態を監視するものである。サーバ制御プロセス124は、BEPノードに必要なサーバプロセスを起動するものである。ユーザI/F処理プロセス125は、システム管理者にユーザインタフェースを提供するものである。IPパケット処理層131は、通信プロトコルの一つであるインターネットプロトコル（IP）のパケットの処理を行うものである。TCPパケット処理層132は、通信プロトコルの一つであるTCP（transmission control protocol）のパケットの処理を行うものである。

【0067】サーバプロセス21は、クライアントからのリクエストを処理するプロセスである。内部ネットワークドライバ22は、FEPノード11との通信を制御するものである。接続監視プロセス23は、クライアントとサーバプロセスとの接続状態を監視するものである。IPパケット処理層241は、インターネットプロトコルのパケットの処理を行うものである。TCPパケット処理層242は、TCPのパケットの処理を行うものである。

【0068】：転送制御表と割当済転送経路一覧表の説明

図6は転送制御表と割当済転送経路一覧表の説明図であり、図6（a）は転送制御表の説明、図6（b）は割当

済転送経路一覧表の説明である。

【0069】図6(a)において、転送制御表112には、サービス識別子、転送先BEPノード識別子、同一サービスに対するリクエストの分配比率である比率が設けてある。

【0070】サービス識別子としてデスティネーション(宛て先)ポートのポート番号を使用し、この例では「8080、8081、8082」が記録されている。転送先BEPノード識別子として、デスティネーションIPアドレスを使用し、この例では例えば、サービス「8080」のデスティネーションIPアドレスには「IP-A、IP-B」が記録され、こおれらの比率として、IP-Aには「1」、IP-Bには「1」と同じ比率が記録されている。

【0071】図6(b)において、割当済転送経路一覧表111には、サービス識別子、クライアント識別子、転送先BEPノード識別子が設けてある。サービス識別子としてデスティネーションポートのポート番号を使用し、この例では「8080、8081、8082」が記録されている。

【0072】クライアント識別子としてソースIPアドレスとソースポート番号の組を使用する。ソースIPアドレスはクライアントのIPアドレスであり、ソースポート番号はプロセス毎に設けられている。この例では、(例えば、図4のクライアント(1))ソースIPアドレス「IP-CL1」には3つのソースポート番号「PORT-CL1-x x 1」、「PORT-CL1-x x 2」、「PORT-CL1-x x 3」が示されている。

【0073】転送先BEPノード識別子として、デスティネーションIPアドレスを使用し、この例では例えば、サービス「8080」のデスティネーションIPアドレスには「IP-A」と「IP-B」が記録されている。

【0074】なお、サービスとポート番号の対応はUNIX OS上のファイルに定義され、クラスタ内の各ノードは、共通のサービスポートを使うものである。

サーバ装置の動作の説明

a: クライアントからリクエストがあった場合の説明
FEPノード11の外部ネットワークドライバ114は、クライアントからのリクエストを受信して、パケットフィルタ113に送る。パケットフィルタ113は、割当済転送経路一覧表111と転送制御表112を用いてパケットを転送するBEPノードを決定し、ノード間結合ネットワーク3を介してパケットをBEPノード12に送る。

【0075】このパケットが転送されたBEPノード12の内部ネットワークドライバ22は、受信したパケットをストリーム処理部に送る。サーバプロセス21には、UNIXが提供するシステムコールインタフェースを用いてTCPパケット処理層242に届いたパケット

を読み出して、リクエストを処理する。

【0076】サーバプロセス21は、リクエストの処理が完了するとクライアントにリプライするために、UNIXのシステムコールを用いてパケットをストリーム処理部に渡す。このリプライのパケットは、内部ネットワークドライバ22、ノード間結合ネットワーク3を介してFEPノード11に転送された後、FEPノード11によって、クライアントに送られる。

【0077】b: ノード故障の場合の説明

10 ノード故障検出プロセス122は、ユーザプロセスとして実現されている。ノード故障検出プロセス122は、定期的に全てのBEPノードが正常動作していることを確認し、BEPノードの故障を検出すると、負荷調整プロセス121を経由してパケットフィルタ113の転送制御表112を書き換えて、故障したBEPノードへのリクエストの分配を停止する。パケットフィルタ113は、割当済転送経路一覧表111内の転送先BEPノードが故障したBEPノードとなっている全てのエントリを削除する。

20 【0078】このBEPノードの故障によって、故障したBEPノードを利用していたクライアントとBEPノード上のサーバプロセスとのコネクションは切断されるが、クライアントがリクエストをリトライすることによって故障は隠蔽される。

【0079】クライアントからの再リクエストはパケットフィルタ113に送られると、パケットフィルタ113は、割当済転送経路一覧表111を調べるが、再リクエストしたクライアントのエントリは既に削除されているため、パケットフィルタ113は、転送制御表112に基づいて新たな転送先を決定してリクエストを転送する。リクエストを転送されたBEPノードでは、通常のリクエストと同様にリクエストを処理してリプライする。

【0080】c: ノード負荷監視の説明

ノード負荷監視プロセス123は、定期的にBEPノード12と通信してBEPノード12の負荷状態を監視する。ノード負荷監視プロセス123は、負荷が高いBEPノード12を発見すると、負荷調整プロセス121に通知し、通知を受けた負荷調整プロセス121は、IOCTLシステムコールを用いてパケットフィルタ113内の転送制御表112を書き換える。

【0081】パケットフィルタ113は、BEPノード12上の接続監視プロセス23にクライアントの接続の状態を問い合わせる。接続監視プロセス23は、IOCTLを使って自ノードのストリーム処理部にコネクション状態を問い合わせ、パケットフィルタ113に通知する。パケットフィルタ113は、未使用のコネクションの転送先だけを変更して転送比率を調整する。

【0082】負荷調整プロセス121は、必要があれば、サーバ制御プロセス124にBEPノード12への

サーバプロセス21の起動あるいは停止を依頼する。このように、本発明によれば、同一サービスに対する複数のクライアントからのリクエストを複数台のサーバノードで負荷分散することが可能となる。また、自動的なサーバノード故障隠蔽機能や、負荷調整機能を持つことにより、システム管理コストを削減することができる。

【0083】

【発明の効果】以上説明したように、本発明によれば次のような効果がある。

(1)：フロントエンドサーバに配置したリクエスト転送制御機構で受信したクライアントからのリクエストをクライアントの識別情報を用いて、同一サービスに対する複数のクライアントからのリクエストを転送するバックエンドサーバを決定してリクエストを転送するため、1種類のサービスの処理を複数台のバックエンドサーバで負荷分散でき、処理性能を向上することができる。

【0084】(2)：リクエスト転送制御機構で、リクエスト転送制御表を用いて転送するリクエストの比率をバックエンドサーバ毎に制御するため、バックエンドサーバの処理能力に応じた負荷分散ができ、より処理性能を向上することができる。

【0085】(3)：ノード故障検出機構がサーバの故障を検出すると、リクエスト転送制御機構で、故障したバックエンドサーバへのリクエスト転送を中止して、クライアントからの再リクエスト時に正常運用している別のバックエンドサーバにリクエストを転送するため、クライアントからの再リクエストにより、バックエンドサーバの故障をクライアントから隠蔽することができる。

【0086】(4)：ノード負荷監視機構が負荷が高いバックエンドサーバを発見すると、リクエスト転送制御機構で、負荷の高いバックエンドサーバへの転送比率を下げて、前記バックエンドサーバ間の負荷を平均化するため、常に負荷を平均化することができ、処理性能を向上することができる。

【0087】(5)：ユーザインタフェース機構で、システム管理者がバックエンドサーバへの転送比率を制御するため、例えば、活性保守のためのバックエンドサーバの転送比率をゼロとしてノードの一時的な切り離しや再組み込みが可能となる。

【0088】(6)：リクエスト転送制御機構は、コネクション状態監視機構にコネクション状態を問い合わせ、サービス中のコネクションに対しては、該サービス終了後に、転送経路の変更を行うため、サービス途中に新たなバックエンドサーバに切り換えることがなく、処理の中断を防止することができる。

【0089】(7)：あるサービスのサーバプロセスが

配置された各バックエンドサーバの負荷が高くなった時に、サーバプロセス制御機構がそのサーバプロセスが未配置なバックエンドサーバにサーバプロセスを起動するため、複数個のバックエンドサーバ全体の負荷を調整することができ、処理性能を向上することができる。

【0090】(8)：リクエスト転送制御機構を、ネットワークの通信制御をするネットワークドライバとパケットの処理をするオペレーティングシステムのパケット処理部の間に配置したパケットフィルタで構成するため、パケットフィルタで1種類のサービスの処理を複数台のバックエンドサーバで負荷分散でき、処理性能が向上する。

【0091】(9)：クライアントの識別情報として、クライアントから受信したリクエストのパケットのソースアドレスとソースポート番号のペアを使用するため、クライアントプロセス毎にリクエストの転送先ノードの制御を行うことができる。

【図面の簡単な説明】

【図1】本発明の原理説明図である。

【図2】実施の形態における装置構成図(1)である。

【図3】実施の形態における転送制御表と割当済転送経路一覧表の説明図である。

【図4】実施の形態におけるクライアント/サーバシステムの説明図である。

【図5】実施の形態における装置構成図(2)である。

【図6】実施の形態における転送制御表と割当済転送経路一覧表の説明図である。

【図7】従来例の説明図(1)である。

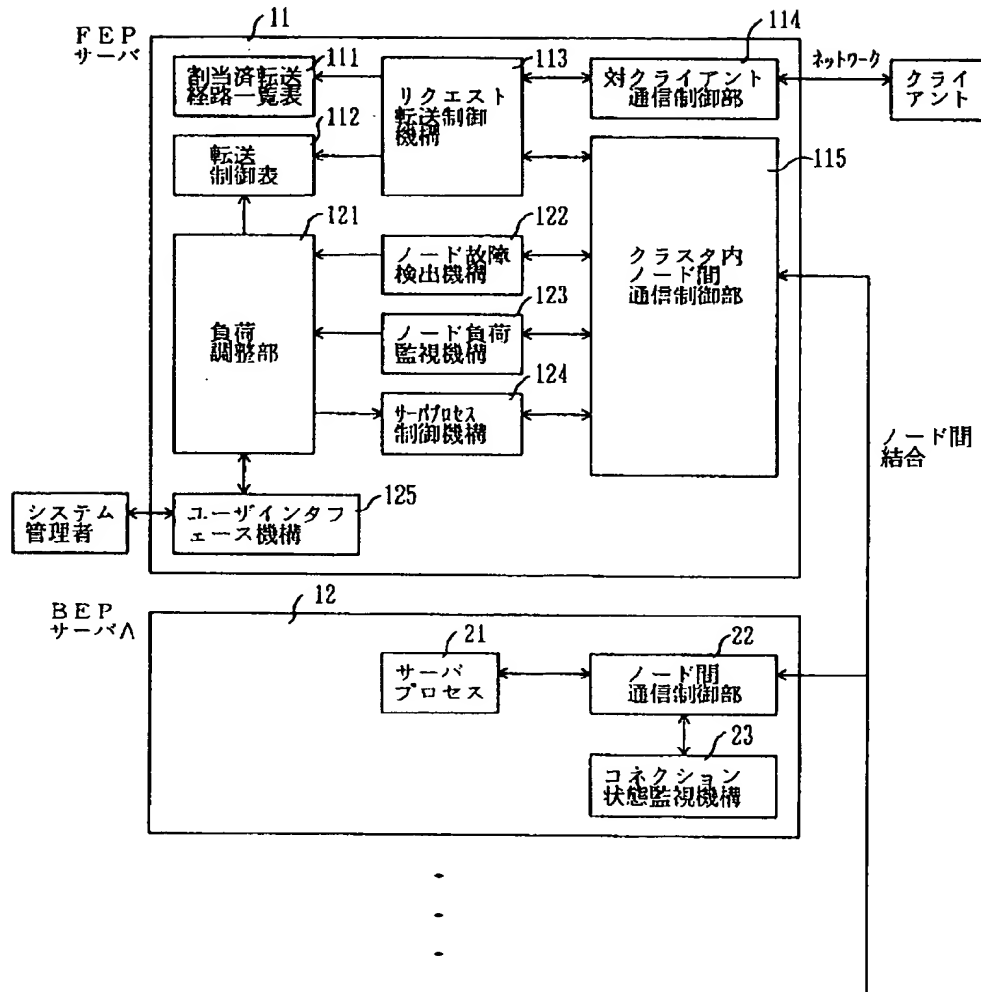
【図8】従来例の説明図(2)である。

【符号の説明】

- 11 フロントエンドサーバ
- 12 バックエンドサーバ
- 21 サーバプロセス
- 22 ノード間通信制御部
- 23 コネクション状態監視機構
- 111 割当済転送経路一覧表
- 112 リクエスト転送制御表
- 113 リクエスト転送制御機構
- 114 対クライアント通信制御部
- 115 クラスタ内ノード間通信制御部
- 121 負荷調整部
- 122 ノード故障検出機構
- 123 ノード負荷監視機構
- 124 サーバプロセス制御機構
- 125 ユーザインタフェース機構

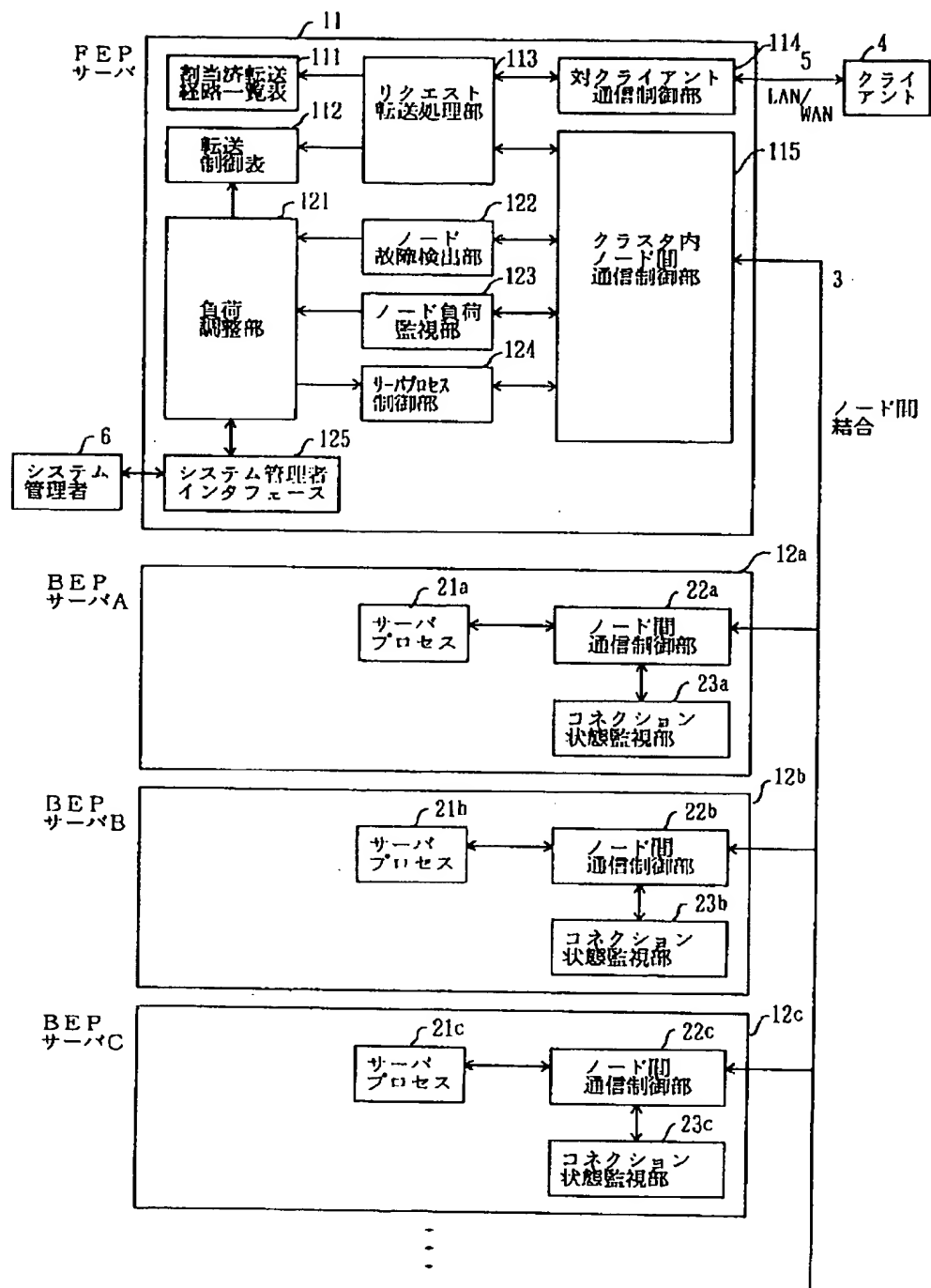
【図1】

本発明の原理説明図



【図2】

装置構成図(1)



【図3】

転送制御表と割当済転送経路一覧表の説明図

(a) 転送制御表の説明

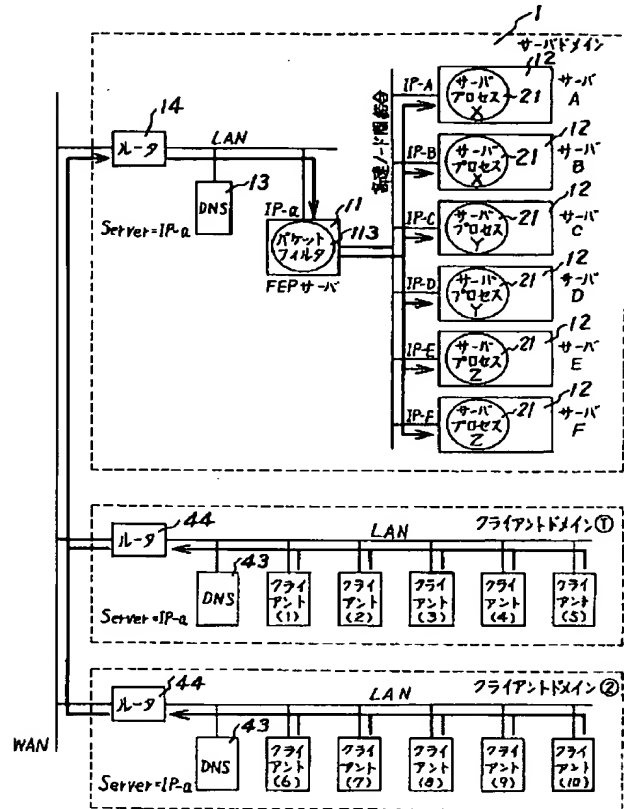
サービス識別子	利用可能BEPノード 識別子	比率

(b) 割当済転送経路一覧表の説明

サービス識別子	クライアント識別子	転送先BEPノード 識別子

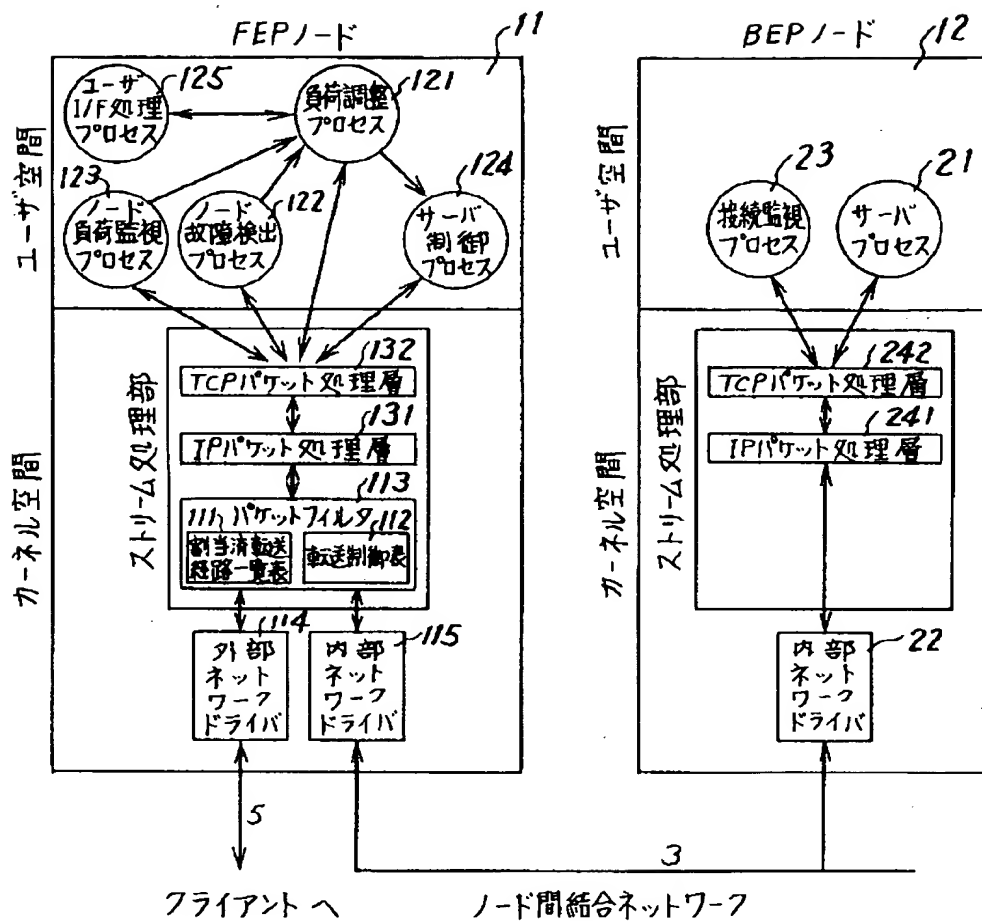
【図4】

クライアント/サーバシステムの説明図



【図5】

装置構成図(2)



【図 6】

転送制御表と割当済転送経路一覧表の説明図

(a) 転送制御表の説明

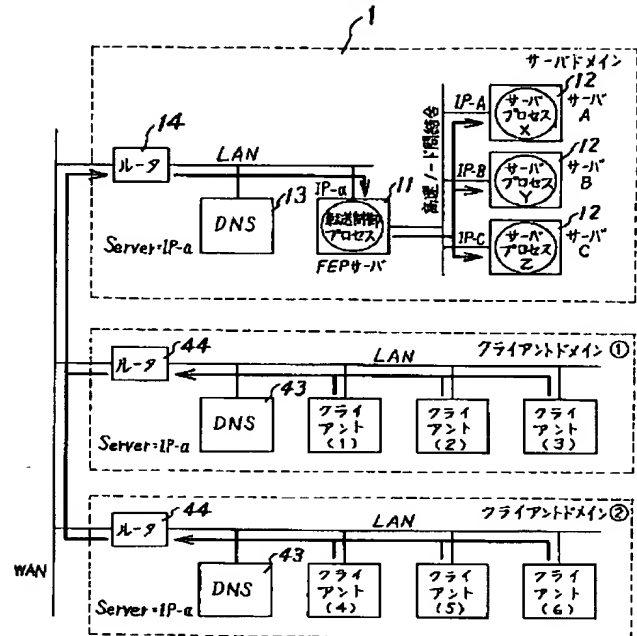
サービス識別子	転送先BEPノード識別子	比 率
デスティネーション ポート番号	デスティネーション IPアドレス	
8080	IP-A	1
	IP-B	1
8081	IP-C	2
	IP-D	1
8082	IP-E	1
	IP-F	1

(b) 割当済転送経路一覧表の説明

サービス識別子	クライアント識別子		転送先BEPノード識別子
デスティネーション ポート番号	ソースIPアドレス	ソースポート番号	デスティネーション IPアドレス
8080	IP-CL1	PORT-CL1-xx1	IP-A
	IP-CL1	PORT-CL1-xx2	IP-B
	IP-CL2	PORT-CL2-xx1	IP-A
	IP-CL3	PORT-CL3-xx1	IP-D
8081	IP-CL4	PORT-CL4-xx1	IP-C
	IP-CL5	PORT-CL5-xx1	IP-D
	IP-CL6	PORT-CL6-xx1	IP-C
	IP-CL7	PORT-CL7-xx1	IP-C
	IP-CL8	PORT-CL8-xx1	IP-D
	IP-CL9	PORT-CL9-xx1	IP-C
8082	IP-CL10	PORT-CL10-xx1	IP-D
	IP-CL1	PORT-CL1-xx3	IP-F
	IP-CL2	PORT-CL2-xx2	IP-B
	IP-CL3	PORT-CL3-xx2	IP-F

【図 7】

従来例の説明図(1)



【図 8】

従来例の説明図(2)

(a) 仕事の割り振り表の説明

サービス識別子	転送先サーバ識別子
サービス X	サーバ A
サービス Y	サーバ B
サービス Z	サーバ C

(b) IPアドレス例の説明

サービス名	IPアドレス
ftp-server	133.160.12.5